PAC Classification based on PAC Estimates of Label Class Distributions *

Nick Palmer
Dept. of Computer Science,
University of Warwick,
Coventry CV4 7AL, U.K.
npalmer@dcs.warwick.ac.uk

Paul Goldberg†
Dept. of Computer Science,
University of Warwick,
Coventry CV4 7AL, U.K.
pwg@dcs.warwick.ac.uk

December 6, 2004

Abstract

A standard approach in pattern classification is to estimate the distributions of the label classes, and then to use the Bayes classifier (applied to the estimates of distributions) to classify unlabelled examples. As one might expect, the better our estimates of the label class distributions, the better will be the resulting classifier. In this paper we verify this observation in the (agnostic) PAC setting, and identify precise bounds on the misclassification rate in terms of the quality of the estimates of the label class distributions, as measured by variation distance or KL-divergence. We show how agnostic PAC learnability relates to estimates of the distributions that have a PAC guarantee on their variation distances from the true distributions, and we express the increase in negative log likelihood risk in terms of PAC bounds on the KL-divergences.

1 Introduction

A common approach to classification problems is to estimate the probability distribution over each label class. An input x is assigned the class label of the class whose associated probability density has the largest value for x (having been weighted by the class prior, if applicable). This required the use of unsupervised learning techniques in order to perform supervised learning. In an earlier paper [5] we study the problems of PAC learning using that general approach, and discuss the advantages the approach has in practice. The aim is to construct the discriminant functions for the k classes in such a way as to minimise the risk of misclassification, rather than to approximate the distribution over inputs. The

*This is an unpublished working document. This work was supported by EPSRC Grant GR/R86188/01. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

†author's home page: `http://www.dcs.warwick.ac.uk/~pwg/`

discriminant function achieving the optimal classification in terms of *risk* (described in Section 1.2) is known as the Bayesian Optimal Classifier [4].

We consider this problem in the context of Probably Approximately Correct (PAC) learning framework [10] and the more general framework of *agnostic* learning [9], in which it is possible for the target concept to be nondeterministic, and not a member of the hypothesis class. The results given here are applicable to learning *probabilistic concepts* (p-concepts)[1], described by Kearns and Schapire [8].

A similar framework is used by Kearns et al. [7], whereby methods for learning a variety of discrete distributions (with respect to variation distance) from stochastic observations are examined. Their framework is not strictly agnostic, as knowledge of the function labeling the data is assumed.

We study two commonly used measures of accuracy - the *variation distance* (or L_1-distance) and the *Kullback-Leibler divergence* (KL-divergence). The KL-divergence is frequently used (for example recently in [2, 6]) due to the property that when minimised with respect to an empirical sample, the likelihood of that sample is maximised. Abe et al. [1] use KL-divergence to learn stochastic rules in a framework similar to the PAC model. A method of avoiding the problem of unbounded log loss is given (ϵ-Bayesian averaging). Their results show that a class of stochastic concepts is learnable in terms of quadratic distance (and therefore variation distance) if and only if it can also be learnt in terms of KL-divergence.

Cryan et al. [3] give an efficient PAC algorithm for learning Markov evolutionary trees using variation distance. In that paper it is shown in a similar way to [1] that if it is possible to PAC-learn a class of distributions over $\{0,1\}^n$ under variation distance then it is PAC-learnable under KL-divergence. This shows that in some cases it is preferable to learn using variation distance rather than KL-divergence.

We show that if it is possible to learn the distribution over label classes within a known accuracy, in the sense described above, then it is possible to agnostically PAC learn the concept. We give specific upper and lower bounds on the risk associated with a hypothesis in terms of both its variation distance and KL-divergence from the target concept.

1.1 Learning Framework

The PAC learning framework dictates that the learning algorithm receives labelled samples generated independently according to distribution D over X, where distribution D is unknown, and where labels are generated by an unknown function f from a known class of functions $\mathcal{F}$. The algorithm must output a hypothesis h from a class of hypotheses $\mathcal{H}$, such that with probability at least $1-\delta$, $err_h \leq \epsilon$, where ϵ and δ are parameters. Notice that in this setting, if $f \in \mathcal{H}$, then $err^* = 0$, where err^* is the error associated with the optimal hypothesis.

The learning scenario described in this paper falls into the agnostic variation of the PAC learning framework. In this framework, nothing is known about the function f which labels the data, therefore the target concept may be stochastic[2] rather than deterministic. Therefore it is the case that $err^* = \min_{h \in \mathcal{H}}\{err_h\}$. The aim of the learning algorithm in this framework is to output a hypothesis $h \in H$ such that with probability of at least $1-\delta$ the error is such that:

[1] p-concepts are functions probabilistically mapping elements of the domain to 2 classes.

[2] In this case, $\mathcal{F}$ is a class of p-concepts as noted by Kearns et al. [9]

$$err_h \leq err^* + \epsilon$$

Our notion of learning distributions is similar to that of Kearns et al. [7].

Definition 1 *Let $\mathcal{D}_n$ be a class of distributions. $\mathcal{D}_n$ is said to be efficiently learnable if an algorithm A exists, such that given $\epsilon > 0$ and $\delta > 0$ and access to randomly drawn examples (see below) from any unknown target distribution $D \in \mathcal{D}_n$, A runs in time polynomial in $\left(\frac{1}{\epsilon}\right)$, $\left(\frac{1}{\delta}\right)$ and n and returns probability distribution h that with probability at least $1 - \delta$ is within ϵ variation distance (alternatively KL-divergence - see Section 1.2) of D.*

We define the model of learning p-concepts as introduced by Kearns and Shapire [8].

Definition 2 *A Probabilistic Concept (or p-concept) on domain X is a real-valued function $c : X \to [0, 1]$. The value $c(x)$ is the conditional probability that sample x is labelled 1. A p-concept class $\mathcal{C}$ is a family of p-concepts. A learning algorithm for $\mathcal{C}$ aims to learn a target p-concept $c \in \mathcal{C}$ with respect to a fixed but unknown and arbitrary target distribution D over X.*

Let function $h : X \to \{0, 1\}$ be a hypothesis (also known as a decision rule). The predictive error of h on c with respect to D, is the probability that h misclassifies a randomly drawn example[3] *(denoted as $R_D(c, h)$). P-concept h is said to be an (ϵ, γ)-good model of probability of c with respect to D if $\Pr_{x \in D}[|h(x) - c(x)| > \gamma] \leq \epsilon$.*

P-concept class C is learnable with a decision rule if there is an algorithm A such that for any target p-concept $c \in \mathcal{C}$, for any target distribution D over X, for any $\epsilon > 0$, $\delta > 0$ and $\gamma > 0$, algorithm A (given access to randomly drawn examples) halts and with probability at least $1 - \delta$ outputs a p-concept h that is an (ϵ, γ)-good model of probability of c with respect to D.

1.2 Notation and Terminology

Given a probability distribution D over elements of $X \times \{1, ..., k\}$, we consider the problem of predicting the label ℓ associated with $x \in X$, where x is generated by the marginal distribution of D on X, $D|_X$. A non-negative penalty is incurred for each classification, based either on a penalty matrix (where the penalty depends upon both the hypothesised label and the true label) or the negative log-likelihood of the true label being assigned. The aim is to optimise the expected penalty given by the occurence of a randomly generated example. We refer to the expected penalty associated with any hypothesis function as *risk* (as described by Vapnik [11]), which we denote as $R(f) = E(pen(f))$.

Let D_ℓ be D restricted to points (x, ℓ) with $\ell = \{1, ..., k\}$, for which the class prior, g_ℓ, is the probability that a randomly generated element has label ℓ.

Let $d_i(x)$ be the difference between the probability densities of D_i and $\hat{D}_i$ at $x \in X$:

$$d_i(x) = |D_i(x) - \hat{D}_i(x)|$$

In Section 2 it is shown that if we know how close the approximated distributions are to the true distributions of each class label, then we can bound the risk associated with

[3] An example is a point $x \in X$ drawn randomly according to D, and then labeled 1 with probability $c(x)$, and 0 with probability $1 - c(x)$.

the classifiers. Suppose D and D' are probability distributions over the same domain X. We define the variation distance as:

$$d_{var}(D, D') = \int_X |D(x) - D'(x)| \, dx$$

For discrete X:

$$d_{var}(D, D') = \sum_{x \in X} |D(x) - D'(x)|$$

The KL-divergence is defined as:

$$I(D||D') = \sum_{x \in X} D(x) \log \left(\frac{D(x)}{D'(x)} \right)$$

We define $I(D||D')(x)$ to be the contribution at $x \in X$ to the KL-divergence. Therefore:

$$I(D||D')(x) = D(x) \log \left(\frac{D(x)}{D'(x)} \right)$$

2 Results

In this section we give bounds on the risk associated with a hypothesis, with respect to the accuracy of the approximation of the underlying distribution generating the instances. We define the accuracy of an approximate distribution in terms of variation distance and KL-distance, both of which are commonly used measurements.

First we examine the case where the accuracy of the hypothesis distribution is such that the distribution for each class label is within variation distance ϵ of the true distribution for that label, for some $0 \leq \epsilon \leq 1$. A penalty matrix A specifies the penalty associated with any classification, where the penalty of classifying a data point which has label i as some label j is denoted as a_{ij} (where $a_{ij} \geq 0$). It is usually the case that $a_{ij} = 0$ for $i = j$.

Given a classification function $f : X \to \{1, ..., k\}$, the risk associated with f is measured by the expected penalty $E(pen(f))$, where (for a discrete domain X):

$$E(pen(f)) = \sum_{x \in X} \sum_{i=1}^{k} a_{if(x)}.g_i.D_i(x)$$

f^* is defined to be the function with the optimal expected penalty, and $\hat{f}(x)$ is the function with optimal expected penalty given alternative distributions $\hat{D}_i, i \in \{1, ..., k\}$.

For $x \in X$:

$$f^*(x) = \arg\min_j \sum_{i=1}^{k} a_{ij}.g_i.D_i(x)$$

$$\hat{f}(x) = \arg\min_j \sum_{i=1}^{k} a_{ij}.g_i.\hat{D}_i(x)$$

Theorem 3 *If for each label* $i \in \{1, ..., k\}$, $d_{var}(D_i, \hat{D}_i) \leq \epsilon$, *then* $R(\hat{f}) \leq R(f^*) + \epsilon . \max_{ij}\{a_{ij}\}$.

Proof: The expected penalty for assigning label j to $x \in X$ is:

$$\sum_{i=1}^{k} a_{ij} . g_i . D_i(x)$$

Let $\tau(x)$ be the increase in expected penalty (increase in risk) for labelling x as ℓ' instead of ℓ:

$$\tau(x) = \sum_{i=1}^{k} a_{i\ell'} . g_i . D_i(x) - \sum_{i=1}^{k} a_{i\ell} . g_i . D_i(x)$$

$$\tau(x) = \sum_{i=1}^{k} (a_{i\ell'} - a_{i\ell}) . g_i . D_i(x)$$

Let $\ell = f^*(x)$ and $\ell' = \hat{f}(x)$. Due to the optimality of f^* on D_i, it must be the case that:

$$\tau(x) \geq 0$$

It is the case that the expected penalty of ℓ' must be less than or equal to that of ℓ on $\hat{D}_i$, given that ℓ' is chosen to be optimal on the $\hat{D}_i$ values:

$$\sum_{i=1}^{k} a_{i\ell'} . g_i . \hat{D}_i(x) \leq \sum_{i=1}^{k} a_{i\ell} . g_i . \hat{D}_i(x)$$

$$\sum_{i=1}^{k} \hat{D}_i(x) . g_i (a_{i\ell} - a_{i\ell'}) \geq 0$$

In order for this to be true, the sum of the difference between the expected penalties generated by ℓ' and ℓ must be greater than $\tau(x)$:

$$\tau(x) \leq \sum_{i=1}^{k} |a_{i\ell'} - a_{i\ell}| . g_i . d_i(x)$$

$$\tau(x) \leq \sum_{i=1}^{k} \max_j\{a_{ij}\} . g_i . d_i(x)$$

In order to bound the expected penalty, it is necessary to sum over the range of $x \in X$:

$$\sum_{x \in X} \tau(x) \leq \sum_{x \in X} \sum_{i=1}^{k} \max_j\{a_{ij}\} . g_i . d_i(x)$$

$\max_j\{a_{ij}\}$ and g_i are both constant and independent of X, so it must be the case that:

$$\sum_{x \in X} \tau(x) \leq \sum_{i=1}^{k} \max_j\{a_{ij}\} . g_i . (\sum_{x \in X} d_i(x))$$

Since it is the case that $\sum_{x \in X} d_i(x) \leq \epsilon$ for all i, it follows that:

$$\sum_{x \in X} \tau(x) \leq \sum_{i=1}^{k} \max_j \{a_{ij}\}.g_i.\epsilon$$

This expression gives an upper bound on expected penalty for labelling x as $\hat{f}(x)$ instead of $f^*(x)$. By definition:

$$\sum_{x \in X} \tau(x) = R(\hat{f}) - R(f^*)$$

Therefore it has been shown that:

$$R(\hat{f}) \leq R(f^*) + \sum_{i=1}^{k} \max_j \{a_{ij}\}.g_i.\epsilon$$

Since $\sum_{i=1}^{k} g_i = 1$, it can be seen that:

$$R(\hat{f}) \leq R(f^*) + \epsilon.\max_{ij}\{a_{ij}\}$$

□

We have given a general upper bound on the increase in risk of $\epsilon.\max_{ij}\{a_{ij}\}$. In some instances, a tighter bound may be given by $R(\hat{f}) \leq R(f^*) + \sum_{i=1}^{k} \max_j \{a_{ij}\}.g_i.\epsilon$, which is also a valid upper bound.

We now prove a similar result in terms of KL-divergence, using the negative log-likelihood of the correct label as the penalty function. Given a function $f : X \to \mathbf{R}^k$, where $f(x)$ is a prediction of the probabilities of x having each label $i \in \{1, ..., k\}$ (so $\sum_{i=1}^{k} f(x)_i = 1$), the risk can be expressed as:

$$R(f) = \sum_{x \in X} D(x) \sum_{i=1}^{k} Pr(label(x) = i).(-\log(f(x)_i))$$

From this equation it can be seen that:

$$R(f^*) = \sum_{x \in X} D(x) \sum_{i=1}^{k} Pr(label(x) = i).(-\log(Pr(label(x) = i)))$$

$$R(f^*) = \sum_{x \in X} D(x) \sum_{i=1}^{k} \left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)} \right) . \left(-\log \left(\frac{D_i(x)}{\sum_{j=1}^{k} \hat{D}_j(x)} \right) \right) \quad (1)$$

Theorem 4 *If for each label $i \in \{1, ..., k\}$, $I(D_i || \hat{D}_i) \leq \epsilon$, then $R(\hat{f}) \leq R(f^*) + k\epsilon$.*

Proof: Let $\tau(x)$ be the contribution at $x \in X$ to the risk associated with $\hat{f}$.

$$R(\hat{f}) = \sum_{x \in X} \tau(x)$$

and

$$\tau(x) = D(x).\sum_{i=1}^{k}\left(\left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)}\right).\left(-\log\left(\frac{\hat{D}_i(x)}{\sum_{j=1}^{k}\hat{D}_j(x)}\right)\right)\right)$$

Let $\xi(x)$ denote the contribution to additional risk incurred from using $\hat{f}$ as opposed to f^* at $x \in X$. From Equation 1 it can be seen that:

$$\begin{aligned}\xi(x) &= \tau(x) - D(x).\sum_{i=1}^{k}\left(\left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)}\right).\left(-\log\left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)}\right)\right)\right)\\ &= D(x).\sum_{i=1}^{k}\left(\left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)}\right).\left(\log\left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)}\right) - \log\left(\frac{\hat{D}_i(x)}{\sum_{j=1}^{k}\hat{D}_j(x)}\right)\right)\right)\end{aligned}$$

Therefore $\xi(x)$ is equal to

$$D(x).\sum_{i=1}^{k}\left(\left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)}\right).\left(\log\left(D_i(x)\right) - \log\left(\sum_{j=1}^{k} D_j(x)\right) - \log\left(\hat{D}_i(x)\right) + \log\left(\sum_{j=1}^{k}\hat{D}_j(x)\right)\right)\right)$$

$$= D(x).\sum_{i=1}^{k}\left(\left(\frac{D_i(x)}{\sum_{j=1}^{k} D_j(x)}\right).\left(\log\left(\frac{D_i(x)}{\hat{D}_i(x)}\right) - \log\left(\frac{\sum_{j=1}^{k} D_j}{\sum_{j=1}^{k}\hat{D}_j}\right)\right)\right)$$

Since it is the case that $D(x) = \sum_{i=1}^{k} D_i(x)$, and similarly that $\hat{D}(x) = \sum_{i=1}^{k}\hat{D}_i(x)$, $\xi(x)$ can be rewritten as:

$$\xi(x) = D(x).\sum_{i=1}^{k}\left(\frac{D_i(x)}{D(x)}\right).\left(\log\left(\frac{D_i(x)}{\hat{D}_i(x)}\right) - \log\left(\frac{D(x)}{\hat{D}(x)}\right)\right)$$

$$= \sum_{i=1}^{k}\left(D_i(x)\log\left(\frac{D_i(x)}{\hat{D}_i(x)}\right)\right) - D(x)\log\left(\frac{D(x)}{\hat{D}(x)}\right)$$

It follows that:

$$\sum_{x\in X}\xi(x) = \sum_{i=1}^{k}\left(I(D_i||\hat{D}_i)\right) - I(D||\hat{D})$$

$$\sum_{x\in X}\xi(x) \leq k\epsilon - I(D||\hat{D})$$

Due to the fact that the KL-distance between two distributions is non-negative, an upper bound on the penalty can be obtained by letting $I(D||\hat{D}) = 0$:

$$R(\hat{f}) - R(f^*) \leq k\epsilon$$

Therefore it has been proved that:

$$R(\hat{f}) \leq R(f^*) + k\epsilon$$

□

2.1 Lower Bounds

In this section we give lower bounds corresponding to the two upper bounds given in Section 2.

Example 5 *Consider a distribution D over domain $X = \{x_0, x_1\}$, from which data is generated with labels 0 and 1 and there is an equal probability of each label being generated. Let $D_i(x)$ denote the probability that a point is generated at $x \in X$ given that it has label i. D_0 and D_1 are distributions over X, such that at $x \in X$, $D(x) = \frac{1}{2}(D_0(x) + D_1(x))$.*

Suppose that $\hat{D}_0$ and $\hat{D}_1$ are approximations of D_0 and D_1, and that $d_{var}(D_0, \hat{D}_0) = \epsilon$ and $d_{var}(D_1.\hat{D}_1) = \epsilon$, where $\epsilon = \epsilon' + \gamma$ (and γ is an arbitrarily small constant).

Given the following distributions, it can be seen that $R(f^*) = \frac{1}{2} - \frac{\epsilon'}{2}$ (assuming that a misclassification results in a penalty of 1, and that a correct classification results in no penalty):

$$D_0(x_0) = \frac{1}{2} + \frac{\epsilon'}{2}, D_0(x_1) = \frac{1}{2} - \frac{\epsilon'}{2}$$
$$D_1(x_0) = \frac{1}{2} - \frac{\epsilon'}{2}, D_1(x_1) = \frac{1}{2} + \frac{\epsilon'}{2}$$

Now if we have approximations $\hat{D}_0$ and $\hat{D}_1$ as shown below, it can be seen that $\hat{f}$ will misclassify for every value of $x \in X$:

$$\hat{D}_0(x_0) = \frac{1}{2}, \hat{D}_0(x_1) = \frac{1}{2} + \gamma$$
$$\hat{D}_1(x_0) = \frac{1}{2} + \gamma, \hat{D}_1(x_1) = \frac{1}{2}$$

This results in $R(\hat{f}) = \frac{1}{2} + \frac{\epsilon'}{2}$. Therefore $R(\hat{f}) = R(f^*) + \epsilon' = R(f^*) + \epsilon - \gamma$.

In this example the risk is only γ under $R(f^*) + \sum_{i=1}^{k} \max_j\{a_{ij}\}.g_i.\epsilon$. A similar example can be used to demonstrate Theorem 4.

Example 6 *Consider distributions D_0, D_1, $\hat{D}_0$ and $\hat{D}_1$ over domain $X = \{x_0, x_1\}$ as defined in Example 5. It can be seen that the KL-divergence between each label's distribution and its approximated distribution is:*

$$I(D_0, \hat{D}_0) = I(D_1, \hat{D}_1) = \left(\frac{1}{2} + \epsilon'\right) \log\left(\frac{\frac{1}{2} + \epsilon'}{\frac{1}{2}}\right) + \left(\frac{1}{2} - \epsilon'\right) \log\left(\frac{\frac{1}{2} - \epsilon'}{\frac{1}{2} + \gamma}\right)$$

The optimal risk, measured in terms of negative log-likelihood, can be expressed as:

$$R(f^*) = -\left(\frac{1}{2} + \epsilon'\right) \log\left(\frac{1}{2} + \epsilon'\right) - \left(\frac{1}{2} - \epsilon'\right) \log\left(\frac{1}{2} - \epsilon'\right)$$

The risk incurred by using $\hat{f}$ as the discriminant function can be seen to be:

$$R(\hat{f}) = -\left(\frac{1}{2} + \epsilon'\right) \log\left(\frac{1}{2}\right) - \left(\frac{1}{2} - \epsilon'\right) \log\left(\frac{1}{2}\right)$$

Therefore it is the case that:

$$R(\hat{f}) = R(^*) + \left(\frac{1}{2} + \epsilon'\right) \log \left(\frac{\frac{1}{2} + \epsilon'}{\frac{1}{2}}\right) + \left(\frac{1}{2} - \epsilon'\right) \log \left(\frac{\frac{1}{2} - \epsilon'}{\frac{1}{2}}\right) = R(^*) + \epsilon - \left(\frac{1}{2} - \epsilon'\right) \log \left(\frac{\frac{1}{2} + \gamma}{\frac{1}{2}}\right)$$

This shows that as γ tends to zero, the increase in risk tends to ϵ.

3 Conclusion

We have shown a close relationship between the error of an estimated input distribution (as measured by variation distance or KL-divergence) and the error rate of the resulting classifier. In situations where we believe that input distributions may be accurately estimated, the resulting information about the data may be more useful than just a near-optimal classifier. Note however that if the estimated input distributions are inaccurate with respect to variation distance, it is still possible for the associated classifier to be of high quality.

References

[1] N. Abe, J. Takeuchi, and M. K. Warmuth. Polynomial learnability of stochastic rules with respect to the KL-divergence and quadratic distance. *IEICE Trans. Inf. & Syst.*, E84-D [3]:299–316, 2001.

[2] A. Clark and F. Thollard. Pac-learnability of probabilistic deterministic finite state automata. *Journal of Machine Learning Research*, 5:473–497, 2003.

[3] M. Cryan, L. A. Goldberg, and P. W. Goldberg. Evolutionary trees can be learnt in polynomial time in the two-state general markov model. *SICOMP*, 31(2):375–397, 2001. copyright SIAM.

[4] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[5] P. W. Goldberg. When can two unsupervised learners achieve PAC separation? In *Procedings of the COLT-EUROCOLT*, number 2111 in LNAI, pages 303–319, 2001.

[6] K. Hoffgen. Learning and robust learning of product distributions. In *ACM COLT*, pages 77–83, 1993.

[7] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions. *Journal of the Association for Computing Machinery*, 1994.

[8] M. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48 [3]:464–497, 1993.

[9] M. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 341–352. ACM Press, 1992.

[10] L. G. Valiant. A theory of the learnable. *Journal of the Association for Computing Machinery*, 27:1134–1142, 1984.

[11] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, second edition, 2000.